

# UNSUPERVISED DECOMPOSITION OF MORPHOLOGY A DISTRIBUTED REPRESENTATION OF THE ITALIAN VERB SYSTEM

*Basilio Calderone*

Laboratorio di Linguistica, Scuola Normale Superiore,  
Piazza dei Cavalieri, 7 - Pisa, ITALY, b.calderone@sns.it

## ABSTRACT

The paper presents a morphological learning process simulated by an ICA (Independent Component Analysis) algorithm. In an unsupervised manner, the algorithm is able to discover emergent morphologically-motivated features from a representative corpus of Italian verbs. The discovered features can be assumed as non-discrete and distributed representations of the morphological data. Final results also reflect a desirable configuration of an associative network, in which different verbal forms share features or categories corresponding to grammatical and paradigmatic attributes. Without reference to the linguistic notion of ‘morpheme’, our system provides a functional model that explains some peculiar aspects of the organization of the mental lexicon.

## 1. INTRODUCTION

Recently in the field of word processing, the nature of the morphological representation of the speaker’s mental lexicon has been a much debated question. Controversies about the role and nature of the morphological information stored in the lexical system, are still open.

Associative models of morphological processes have been developed to explain the relation of form and meaning within interconnected forms [1]. In this framework, words are fully stored in the speaker’s lexicon, including complex words and irregular inflected forms [2]. According to these models, the frequency of words sets the associative strength for all the sets of forms in lexical storage, thereby guaranteeing processes for both the productivity and retrieval of morphological data [3]. Thus all of the morphologically defined structures are the outcome of a processing of a single full-form that is connected to other words on the basis of their orthographic and semantic properties. Some other studies, instead, have shown the coexistence of the associative model with morphological structures decomposed (as affixes and stems) to process the words [4]. This approach, in its versions [5], claims that full-form representations must be combined with the morpheme-based decompositions to guarantee the correct processing of data, including irregular inflected forms (generally morpheme-based decomposition is provided for regular forms, whereas the irregular forms are fully listed in the speaker’s lexicon. See [5] for an overview). Other studies outline different properties as important structur-

ing elements of the mental lexicon. Morphosyntactic properties seem to play a crucial role in experiments of free recall tasks [6]. In particular, with reference to Italian verbs, information about conjugation and mood is relevant to the mental organization of the verb system [7]. On the other hand, psycholinguistic and linguistic evidences [8], support the hypothesis that paradigmatic relations exert a polarizing action on the stored forms, thus defining processes of comparison and opposition within the paradigmatic dimension. Also, the theoretical notion of ‘morpheme’ as the minimal part of form and meaning for compositional processes [9] has to be reconsidered. Many authors have claimed that morphological processing is performed via full-form representations [2, 10]. They also state that it is preferable to deal with morphological markers by treating them as morphosyntactic attributes, rather than by treating them as morphemic units [11]. Recent studies, in a probabilistic perspective, suggest that the morphological structures are inherently graded [12]. In particular the morphological gradience emerges as the result of the probability distribution of distinct forms (in full-word representations) combined with analogy-based processing (as, for example, the paradigmatic analogy [13]). Rejecting the classical-localist approach of the ‘morphemic representation’, this view considers the emergence of non-discrete morphological representations as a consequence of statistical regularities presented by data during the learning phases. In this sense, the morphology is guided by gradient structures modelling the morphology according to solely principles of the data as the frequency, the statistical regularities, the paradigmatic analogy, without any exogenous parameters of supervision.

In this paper we report a simulation of an unsupervised learning of the Italian verb system, using an ICA algorithm, with the purpose of finding non-discreted (morphological) representations guiding the belonging of the morphological data to well-defined linguistic categories.

## 2. UNSUPERVISED METHODOLOGIES

Unsupervised learning regards the capability to discover of structural relevances from data that are initially raw and not structured. An essential property of unsupervised learning systems is that no pre-model or external parameters are defined *a priori*: the learner (after training) proposes by itself a suitable model (with its values and pa-

rameters) that explains data which have been previously learned. Considering morphological learning as an unsupervised learning, presupposes that raw data have underlying morphological structures; in other words, it means that simple data input already contains all the information needed to guarantee correct linguistic performances such as, for example, the morphological segmentation. Recent unsupervised learning systems [14, 15], that use only raw text as input, have been implemented to output morphologically-based segmentation performances. These approaches develop a morphological grammar which is based on a set of heuristics. Those are to be found, accepted and adopted by the system to maximize the *descriptive efficiency* of all morphemes necessary for the correct analysis of the words. These systems use primitive types such as stems, suffixes and signatures (see [14] for the definition of *signature*). Their major problem is that the above mentioned approaches are too linguistically-oriented, meaning that they assume some linguistic knowledge on how the data should be processed. The segmentation is considered as a sequential processing, aimed at locating morphological components in the words. In our view, instead, no additional information about the means and methodology for processing is to be supplied to the system, apart from the data input. The use of Independent Component Analysis (ICA) aims to provide an unsupervised learning *scenario* for morphological processing (see [16] for ). We have applied an ICA algorithm to a representative corpus of Italian verb system in order to extract a number of morphological features later recognized as linguistic categories.

### 3. INDEPENDENT COMPONENT ANALYSIS

Independent Component Analysis (ICA) [17] is a statistical and computational technique that is able to find latent factors underlying a set of multivariate observations. Specifically, ICA carries out feature extraction from a set of measured data or signals by assuming that observed data are a linear combination of unknown hidden variables statistically independent and nongaussian. In an ICA framework a mixed model is:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad (1)$$

where  $\mathbf{x} = (x^1, x^2, \dots, x^n)^T$  is the vector of observed variables,  $\mathbf{s} = (s^1, s^2, \dots, s^n)^T$  is a vector of variables called independent components and  $\mathbf{A}$  is a mixing matrix. This equation can be inverted and expressed as follows:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \quad (2)$$

where the weighting matrix  $\mathbf{W}$  equals the inverted mixing matrix  $\mathbf{A}$ . One independent component can be expressed by the following equation:

$$s_i = \mathbf{w}_i^T \mathbf{x} = \sum_j w_{ij} x_j. \quad (3)$$

Here we should note that only the mixed data  $\mathbf{x}$  are observable, whereas the source components  $\mathbf{s}$  and the mixing matrix  $\mathbf{A}$  are not. ICA finds  $\mathbf{A}$  and  $\mathbf{s}$  by observing solely

$\mathbf{x}$ . In other words, ICA determines the decomposition of (1) in an unsupervised manner.

In the following section we give further details on ICA methodology.

#### 3.1. ICA: details

The ICA method presupposes some conditions for its performance:

- *Statistically independent components.* As its name suggests, ICA needs signal sources to be statistically independent (independent components). Two components are independent if any knowledge about one implies nothing about the other. This assumption, of course, does not affect the observed data  $\mathbf{x}$ .
- *Nongaussian components.* Strictly connected to statistical independence of components condition, ICA assumes that component have a nongaussian distribution [17]. Therefore maximizing nongaussianity of  $\mathbf{w}^T \mathbf{x}$  in (3) is a guiding point to estimate  $\mathbf{s}$ <sup>1</sup>.
- *Indeterminable variances of components.* Variances of  $\mathbf{s}$  cannot be determined. This assumption is a consequence that both  $\mathbf{A}$  and  $\mathbf{s}$  are unknown. Any scalar multiplier in one of the sources,  $s_i$ , can always be cancelled by dividing the corresponding columns of  $\mathbf{A}$ ,  $a_i$ , by it. In this sense, we fix magnitudes of independent components by assuming unit variance:  $E\{s_i^2\} = 1$ . There remains only an ambiguity of sign: one could multiply a component by  $-1$  without affecting the model.
- *Indeterminable order of components.* Since  $\mathbf{A}$  and  $\mathbf{s}$  are unknown, the order of components cannot be determined. The order of  $\mathbf{A}$  and  $\mathbf{s}$  can be exchanged and it is possible to call any of the independent components as the first one.

#### 3.2. ICA and Morphology

The capability to reveal underlying features, together with unsupervised data processing, makes ICA methodology a plausible tool for simulating a morphological learning process. A possible implementation of ICA in morphological analysis requires to deal with morphological data as complex structures. In an ICA generative model (Eq. 1), morphologically-defined forms can be considered as mixed variables  $\mathbf{x}$  of a (linear) combination of different components. Each of these components contributes, with different efficacy, to the final realization of  $\mathbf{x}$ . One morphological well-formed form (like a verb), in our terms, has to be intended as an observed variable  $\mathbf{x}$ , i.e. as the result of the composition action carried on by the mixing matrix  $\mathbf{A}$  and by the independent components  $\mathbf{s}$ . These independent components represent ‘primitive ingredients’

<sup>1</sup>Maximizing nongaussianity of  $\mathbf{w}^T \mathbf{x}$  is just what the ICA algorithm, used in our work, does. For further information on this algorithm, called FastICA, see <http://www.cis.hut.fi/projects/ica/fastica/> [18].

of the *datum* itself, and can be interpreted as linguistically-motivated features, or markers, used to specify the grammatical and lexical properties of the *datum*. The focus of this work is to investigate how it is possible for hidden features of linguistic categories to emerge during an unsupervised learning of morphologically encoded data. The emergence of linguistic underlying features, extracted from a morphologically-defined set of data, has to take into account also frequency effects in the distribution of input data. Obviously type and token frequency effects may encourage different representational aspects of morphological data.

We believe that the statistical independence condition, adopted by ICA, confers this methodology more computational interest than that obtained by classical methods, such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) [19] which are based on the uncorrelatedness of components (for a comparison see [20]). The independence of the components implies the uncorrelatedness, while the uncorrelatedness does not imply the independence of the components. In order to demonstrate that, we need to define the notion of *covariance*. We know that the covariance is the measure of how much two random variables vary together. The covariance  $Cov$  between two real random variables  $X, Y$  in a probability space  $(\Omega, \mathcal{P}(\Omega), P)$ , where  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ ,  $\mathcal{P}(\Omega)$  denotes the subparts of  $\Omega$  and  $P$  is a probability measure, is defined by

$$Cov(X, Y) := \sum_{i=1}^n P(\omega_i)X(\omega_i)Y(\omega_i) - \left( \sum_{i=1}^n P(\omega_i)X(\omega_i) \right) \left( \sum_{i=1}^n P(\omega_i)Y(\omega_i) \right)$$

In a practical instance let us assume  $\Omega = \{\omega_1, \omega_2, \omega_3\}$  with uniform probability  $P(\omega_i) = 1/3, i = 1, 2, 3$ . In this space let us define two real random variables  $X$  and  $Y$  by

$$X(\omega_1) = 1, X(\omega_2) = 0, X(\omega_3) = -1,$$

$$Y(\omega_1) = 0, Y(\omega_2) = 1, Y(\omega_3) = 0;$$

then it is easy to prove that, although  $Cov(X, Y) = 0$  (that means that the two variables are uncorrelated), the two variable  $X$  and  $Y$  are not independent, because

$$X = 0 \implies Y = 1,$$

which means that any information on  $X$  provides information about  $Y$ .

The main success of ICA lies on the assumption that statistically independent variables can generate mixtures of signals that represent the observed data. ICA can decompose the observed data and recover the original (independent) sources.

The following sections describe the decomposition of Italian verbs, obtained by ICA algorithm according to (Eq. 1), to discover plausible linguistically-motivated features affecting a mental lexicon representation of morphological data. Sensitivity to frequency effects will also be investigated by means of the ratio between input (type/token) frequency and number of features extracted. The inflectional richness of Italian conjugation is a good benchmark to test our hypothesis.

#### 4. THE ITALIAN CONJUGATION

Morphology of Italian verb is characterized by a complex inflectional paradigm with a consistent number of subsets to account for irregular and sub-regular verbs [21]. All of the Italian paradigmatic cells are filled by means of a basic stems/suffixes concatenation process. Inflection suffixes encode morphosyntactic information on tense, mood, person, number and conjugation. As reported in descriptive grammars [22], in the Italian verb system we distinguish three verbal classes corresponding to the three different conjugations on the base of the thematic vowel present in the infinitive form, between the stem and the inflectional suffix. First conjugation, with the thematic vowel **-a**, is the most productive class. With the highest relative distribution (72.4% of the Italian verb classes [23]) it largely covers the set of regular verbs. Neologisms and foreign loan words all fall into it. The second conjugation, with the thematic vowel **-e**, includes mostly irregular verbs (its distribution is 16%) and the third conjugation, defined by **-i** as thematic vowel, has a distribution equal to 11.6% and is composed mostly by regular and partially productive verbs. Besides the conjugation level, also paradigms seem to define highly natural inflectional classes. Different approaches [21] revalue the paradigmatic dimension for controlling the stem alternation function. These studies show, by means of different computational evidence, that the form of stems is directly determined by the information that defines each paradigmatic cell. Paradigmatic information, therefore, forces the entire verb system in several micro-classes that appear to be determinant in morphological learning [10] and may have exerted a convergent pressure in the history of Italian verb system. One Italian verb looks like a complex linguistic object, determined by lexical and morphosyntactic information, as well as a paradigmatic action forcing verb into sub-classes.

#### 5. UNSUPERVISED LEARNING BY ICA

As previously said, our goal is to obtain a decomposition process in (Eq. 1) by using an ICA algorithm, with  $\mathbf{x}$  as morphologically encoded Italian verbs. In the following paragraphs, we provide more details of the model design and input data used for learning.

##### 5.1. Input data

Our input data are verbs written in standard Italian orthography. The verb corpus includes all the simple verbal

forms, about 51, of 30 inflectional paradigms. According to the relative distribution of the Italian verb conjugation [23] mentioned above, 19 of these belong to the first conjugation, 6 paradigms are picked up from the second conjugation and 5 from the third. We consider the single verb segments as rows of  $\mathbf{X}$ . A binary encoding of 30 components (since we define 30 distinct segments in our Italian morphological inventory) is adopted to assign an orthogonal vector for each morphological segment<sup>2</sup>. In details, every segment in  $\mathbf{X}$  is, in a binary fashion, coded together with its left and right morphological context. This methodology allows an individual segment (in *Focus* position) to be computed with its complete morphotactic environment (*Left/Right* context) during the creation of the matrix  $\mathbf{X}$ . In other terms, we create a morphological context matrix  $\mathbf{X}$  in which  $x_{ij}$  marks the  $i$ th morphological segment in the  $j$ th morphological context, latter is arranged in a focus and left/right context sizes (see Figure 1). The number of the columns in the matrix  $\mathbf{X}$  is defined by the longest word of the corpus. It is worth noting that our ‘windowing’ approach ensures a full-form encoding, without any alignment adjustments or adopting of *ad hoc* solutions.

	Left Context	Focus	Right Context
...	...	$x_{ij}$	...
V	- - - -	V	O L A R E
O	- - - -	V	O L A R E -
L	- - -	V O	L A R E - -
A	- -	V O L	A R E - - -
R	-	V O L A	R E - - - -
E	V O L A R E	E	- - - - -
...	...	$x_{ij}$	...

Figure 1. Morphological encoding for creating the matrix  $\mathbf{X}$ . The example shows encoding for *volare*, ‘to fly’. The rows of  $\mathbf{X}$  are morphological segments of verbal forms which are defined on the basis of columns of  $\mathbf{X}$ , specifying contextual (morphological) information. The number of the columns in the matrix  $\mathbf{X}$  is defined by the longest word of the corpus. A binary encoding of 30 components is used to orthogonally represent each segment in  $\mathbf{X}$ .

## 5.2. Model design

Figure 2 illustrates the adaptation of the generative model in (1) for our purposes. Matrix  $\mathbf{X}$  is meant as a linear combination of linguistically-motivated features  $\mathbf{S}$  according to the values of a mixing matrix  $\mathbf{A}$ .

Our attention will be focused just on  $\mathbf{A}$ . Each column in  $\mathbf{A}$  corresponds to one single feature that contributes, multiplied by  $\mathbf{S}$ , to reconstructing  $\mathbf{X}$ , i.e. the input morphological segments. In this view, all the rows of the mixing matrix  $\mathbf{A}$  compose a  $N$ -dimensional<sup>3</sup> vector representation for each morphological segments.

<sup>2</sup>See [24] for ICA applied to binary data.

<sup>3</sup>For reasons of computational performance, we reduce the matrix  $\mathbf{X}$

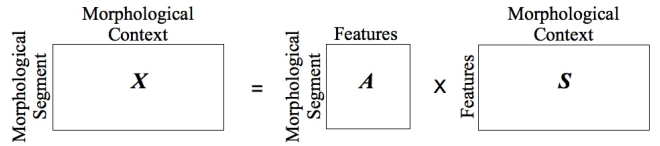


Figure 2. The ICA model adopted for linguistic (morphological) features extraction from matrix  $\mathbf{X}$  using mixing matrix  $\mathbf{A}$  and independent components  $\mathbf{S}$ .

## 5.3. Results

After ICA application to  $\mathbf{X}$ , the mixing matrix  $\mathbf{A}$  is investigated. The ICA algorithm was applied setting 50 independent components for the decomposition (1) of the morphological data matrix,  $\mathbf{X}$ , thus  $\mathbf{A} = a_{ij}$  where  $j = 1, \dots, 50$  whereas the index  $i$  marks the  $i$ th morphological segment. The results are satisfactory.

Emergent linguistically-motivated features are discovered by the system and defined as markers of morphological information. Verbal data are disassembled in different distributed features matching with linguistically-known categories. In particular, after a statistical distribution analysis, emergent features correspond to grammatical and paradigmatic attributes. The first ones provide information on morphosyntactic aspects as tense, mood, person and number. Paradigmatic attributes, instead, mark the paradigmatic family with which the verb (and its lexical class) is related. Inside the corpus, we can distinguish a typology of  $\mathbf{G}$ , grammatical or morphosyntactic features, and of  $\mathbf{P}$  attributes, paradigmatic information features. Verbs of the same grammatical classes, share the same  $\mathbf{G}$ -type attributes. On the contrary, verbs of the same paradigmatic classes have  $\mathbf{P}$ -type attributes in common. Due to the large corpus, we present a limited number of example of features extraction. Figure 3 reports the results for four  $\mathbf{G}$ -type attributes recognizable as morphosyntactic features.

We plotted the segments (one segment for each verbal form) that maximize the selected component. In other terms, about 1500 verbal segments are plotted against different dimensions (among the 50 dimensions) of the mixing matrix  $\mathbf{A}$ . The segments are expressed as cross-markers. Thus, the y-axis defines the value of one selected feature for the word segments and the x-axis indicates the number of the word segments. Circled crosses define the verb segments belonging to a particular grammatical category, i.e. defined by morphosyntactic attributes. It may easily be noted that high values of y-axis (selected component) mostly refer to segments that are grammatically marked. The 10th component of  $\mathbf{A}$  is significant for a morphosyntactic attribute like FIRST PERSON PLURAL, whereas the IMPERFECT INDICATIVE<sup>4</sup> is marked by 5th component of  $\mathbf{A}$ . The ICA algorithm individuates hid-

to 50 components by using Principal Component Analysis. This reduction is a procedure of the FastICA algorithm we have used [18].

<sup>4</sup>The Imperfect tense describes a situation in the past, namely an event which was ongoing or repeated. In English it is often expressed by past continuous tense.

den representations of (statistically) implicit structure in the input  $\mathbf{X}$  matching plausibly grammatically-based attributes. Paradigmatic attributes (**P**-type) examples are displayed in Figure 4. Squared-crosses indicates verb segments paradigmatically-related to the same reference lexeme. Inflected forms of the paradigmatic structure DEFINIRE ‘define’ (squared-cross markers) are marked out for high weights of the 41<sup>st</sup> component of  $\mathbf{A}$ . The 48<sup>th</sup> component, instead, specify the verbal forms of the paradigm AMARE ‘love’.

### 5.3.1. Distributed representations

The features (or attributes) extracted by ICA algorithm can be assumed as  $N$ -dimensional representations marking morphological information. In this sense we say that they provide emergent distributed and non-discrete representations of verbal data. In such distributed representation each form is represented by vectors of attributes, i.e. the independent components, that cluster morphological data by means of fragmented (latent) linguistic information and not by single-valued discrete variables (as the ‘morphemic representation’). The emergent attributes can account for morphological similarity and, at the same time, determine degrees of linguistic category membership: the degree of membership of category, for a verb, is specified by the its number of features specifying the categories (in our terms, grammatical and paradigmatic categories). See Figure 5 for an illustration of the distributed morphological representations of verbal forms. A more detailed analysis of the number of features that determine category membership is likely to highlight micro-classes of prototypical morphological stem patterns, ‘islands of reliability’ [25], that are high productive in selecting specific subregular suffixes.

In an associative framework [2], the **G** and **P**-type features can be arranged in one grid of an associative model in which morphosyntactic and paradigmatic information are related to each other, in order to guarantee the efficient processing for retrieval, interpretation and productivity of new verbal forms. In our framework, an associative network of morphological processing can be designed by ‘setting in rows’ the decomposed morphological segments to reconstruct the order of segments for each verb.

Four verbal forms, picked up from the corpus, are presented in Figure 5. The same morphosyntactic attributes are shared by the forms belonging to the same grammatical classes. The attributes in common for *dicevo* ‘I was saying’ and *cammino* ‘I was walking’ are defined by **G**(12), **G**(21), **G**(14), **G**(22), **G**(6), **G**(5), **G**(17) and **G**(37). These **G**-type attributes plausibly specify the class 1 SINGULAR IMPERFECT INDICATIVE. Other grammatical classes, defined by **G**(19), **G**(9), **G**(8), **G**(18) and **G**(4), are traced in the representations of *dicevate* ‘You were saying’ and *camminate* ‘You walk’, both as 2<sup>nd</sup> person plural. These emergent features are likely to mark the class 2<sup>nd</sup> PLURAL. An interesting aspect is given by the features **G**(12), **G**(14) and **G**(5) that are shared by *cammino*, *dicevo* and *dicevate* ‘You were saying’. The grammatical class IMPERFECT INDICATIVE can be expli-

cated by these common components. Paradigmatic relations are also expressed in our associative network. In Figure 5, verbal forms of the same paradigmatic structure as *cammino* and *camminate* (or *dicevo* and *dicevate*) are marked by **P**-type attributes indicating paradigmatic relations among the verbs of same paradigmatic family. A frequency sensitivity is also exhibited by the system. To test this evidence a new decomposition by ICA was implemented by sampling the corpus verbal forms according to their probability densities in a free context corpus of 5000 words. High-frequency forms effects are accounted for by means of the number and of the weight (i.e. the ‘intensity value’) of the features discovered. The number of grammatical features, extracted by ICA algorithm, seems to be incrementally related to the frequency of the grammatical forms. High-frequency of forms belonging to grammatical classes (as the IMPERFECT INDICATIVE) are defined by features (for **G**-type features), whose number increases until an equilibrium point is reached. This is because the features gain the verb until they reach a critical point of stabilization.

This descriptive compromise regards the whole set of features (**G** and **P**-type) and it is yielded by the ratio between the number of features and the data they have to represent. Another interesting aspect of frequency role is given from *tokens* frequently attested in the corpus. The weights of the features, i.e. the independent components, have a higher prominence for frequent verb forms, compared with the normal average that is calculated in frequency normalized conditions.

## 6. FUTURE DIRECTIONS

We have reported an application of the ICA algorithm to Italian verb system. The goal of the experiment was to obtain an extraction, in an unsupervised fashion, of linguistically-motivated features from orthographically encoded input data. Our final results look promising. Underlying grammatical and paradigmatic information features are detected by the system and considered as independent components of the verb corpus. The emergent features provide also a distributed representation of the morphological data. The distributed representations of morphological information can be interpreted as a low-dimensional reduction process of morphological data that drives a possible order in the mental lexicon. Due to ‘windowing’ encoding of the input data, the decomposition of ICA is applied to the full-form without alignment’s adjustments. In our opinion this aspect sounds interesting, since micro-class paradigmatic relations, crucial for morphology learning, can be investigated in detail [25]. The decomposition processing can be also guided by frequency effects. *Token* and *class token* frequency specify respectively the intensity value and the total number of the discovered features.

Different future modifications are desirable:

- Adopting a morphological encoding for input data. Orthography level is biased from a psycholinguistic

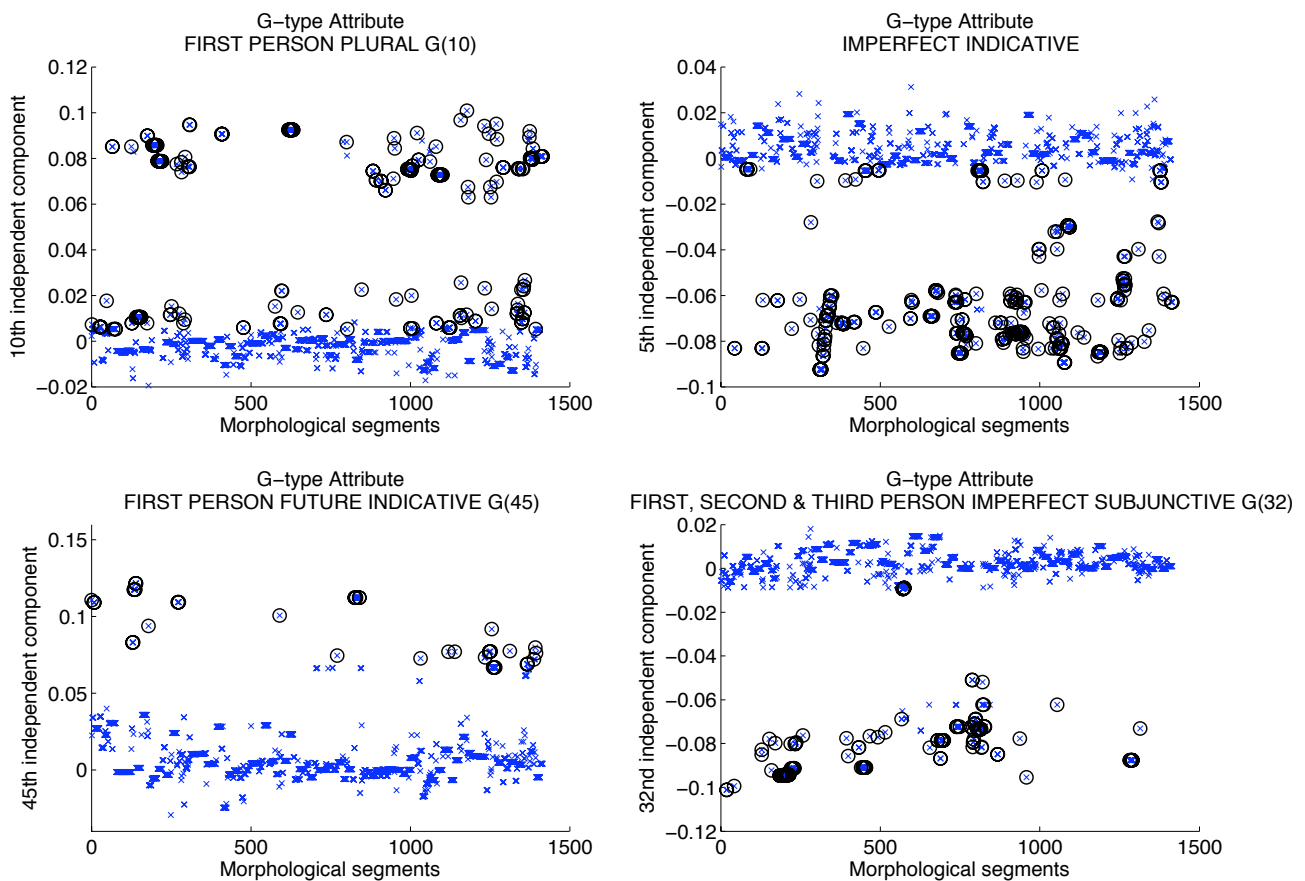


Figure 3. Plotting of the segments, one for each corpus verbal form, maximizing the selected component. Examples of **G**-type attributes, 10<sup>th</sup>, 5<sup>th</sup>, 45<sup>th</sup>, and 23<sup>rd</sup> component of **A**. High values of these components in y-axis specify quite clearly the belonging of the verb segments (cross markers) to different grammatical classes (circled-cross markers), in terms of morphosyntactic attributes (as IMPERFECT INDICATIVE).

point of view. More phonologically-inspired input representations need to be considered.

- Increasing the number of independent components for the extraction of features. We surmise that a deeper specification of features should provide for more grained morphological representations.
- Obtaining a fixed  $N$ -dimensional vector representation of features for each verbal forms by overlapping the distributed representations of morphological verbal segments. For this purpose, aspects of descriptive economy have to be analyzed in detail to highlight the critical compromise between the features' number and the data they have to represent.

## 7. REFERENCES

- [1] M. Raveh and J.G. Rueckl, "Equivalent effects of inflected and derived primes: Long-term morphological priming in fragment completion and lexical decision," *Journal of Memory and Language*, vol. 42, pp. 103–119, 2000.
- [2] J. L. Bybee, "Regular morphology and the lexicon," *Language and Cognitive Processes*, vol. 10, pp. 425–455, 1995.
- [3] J. L. Bybee and P. Hopper, Eds., *Frequency and the emergence of linguistic structure*, John Benjamins, Amsterdam, 2000.
- [4] T. Say and H. Clahsen, "Words, rules and stems in the Italian mental lexicon," in *Storage and computation in the language faculty*, S. Nootboom, F. Weerman, and F. Wijnen, Eds. Kluwer, Dordrecht, 2001.
- [5] H. Clahsen, "Lexical entries and rules of language: a multi-disciplinary study of German inflection," *Behavioral and Brain Sciences*, vol. 22, pp. 991–1060, 1999.
- [6] B.C. Rapp and A. Caramazza, "The modality-specific organization of grammatical categories: Evidence from impaired spoken and written sentence production," *Brain and Language*, vol. 56, pp. 248–286, 1977.
- [7] A. Laudanna, S. Gazzellini, and M. De Martino, "Representation of grammatical properties of Italian

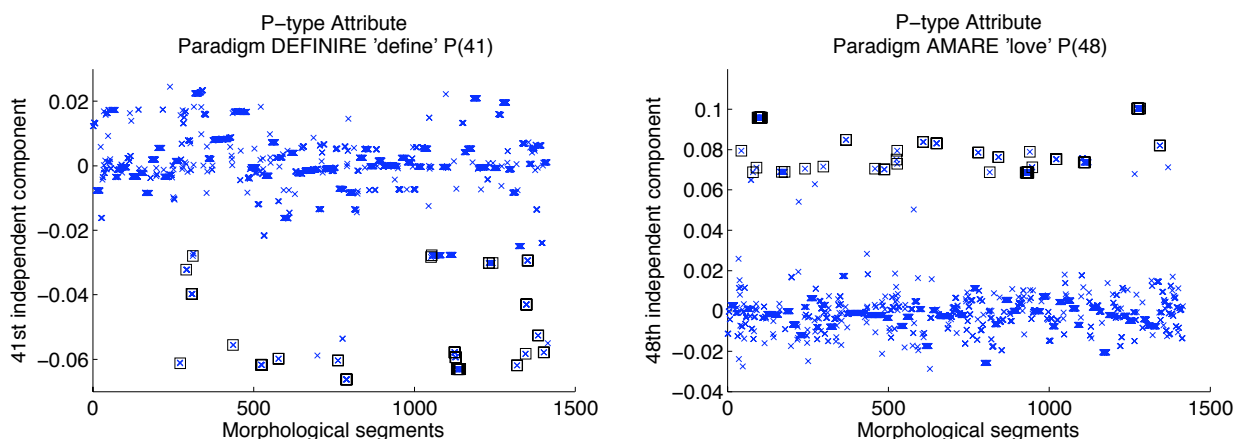


Figure 4. Plotting of the segments, one for each verbal form, maximizing the selected component. Examples of **P**-type attributes, 41<sup>st</sup> and 48<sup>th</sup> component of **A**. High values of the two components (in y-axis) define mostly a distinctive categorization of all the inflected forms of the paradigm (squared-cross markers) of, respectively, DEFINIRE ‘define’ and AMARE ‘love’. It is worth remembering that, for the interpretation of ICA, the value of each component (positive or negative) is irrelevant (see Section 3.1).

- verbs in the mental lexicon,” *Brain and Language*, vol. 90, pp. 95–105, 2004.
- [8] M. Penke, U. Janssen, and S. Eisenbeiss, “Psycholinguistic evidence for the underspecification of morphosyntactic features,” *Brain and Language*, vol. 90, pp. 423–433, 2004.
- [9] R. Lieber, *Deconstructing Morphology*, Chicago University Press, Chicago, 1992.
- [10] J. Blevins, “Stems and paradigms,” *Language*, vol. 79, no. 4, pp. 79(4):737–767.79(4):737–767.Laudanna, Alessandro, Simone Gazzellini, and Maria De 737–767, 2003.
- [11] S. R. Anderson, *A-morphous morphology*, Cambridge University Press, Cambridge, 1992.
- [12] J. Hay and H. Baayen, “Shifting paradigms: gradient structure in morphology,” *Trends in Cognitive Science*, vol. 9, no. 7, pp. 342–348, 2005.
- [13] A. Krott, P. Hagoort, and H. Baayen, “Sublexical units and supralephical combinatorics in the processing of interfixed dutch compounds,” *Language and Cognitive Processes*, vol. 19, no. 3, pp. 453–471, 2004.
- [14] J. Goldsmith, “Unsupervised learning of the morphology of a natural language,” *Computational Linguistics*, vol. 27, no. 2, pp. 153–198, 2001.
- [15] M. Creutz and K. Lagus, “Induction of a simple morphology for highly-inflecting languages,” in *Proceedings of the 7th Meeting of the ACL SIGPHON*, Barcelona, Jul. 2004, pp. 43–51.
- [16] K. Lagus, M. Creutz, and S. Virpioja, “Latent linguistic codes for morphemes using independent component analysis,” in *Proceedings of the 9th Neural Computation and Psychology Workshop (NCPW9)*, A. Cangelosi et al., Ed., Plymouth, England, 2005, pp. 129–138.
- [17] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, J. Wiley, New York, 2001.
- [18] A. Hyvärinen, “Fast and robust fixed-point algorithms for independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–626–634, 1999.
- [19] H. Abdi, “Singular value decomposition and generalized singular value decomposition,” in *Encyclopedia of Measurement and Statistics*, Neil Salkind, Ed., pp. 907–912. Thousand Oaks (CA): Sage, 2007.
- [20] J. Väyrynen and T. Honkela, “Comparison of independent component analysis and singular value decomposition in word context analysis,” in *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation (AKRR)*, Honkela, Könönen, Pöllä, and Simula, Eds., pp. 135–140. Otamedia Oy, 2005.
- [21] V. Pirrelli and M. Battista, “The paradigmatic dimension of stem allomorphy in italian inflection,” *Italian Journal of Linguistics*, vol. 12, no. 2, pp. 307–380, 2000.
- [22] L. Serianni, *Grammatica italiana: italiano comune e lingua letteraria*, UTET, Turin, 1988.
- [23] T. De Mauro, F. Mancini, M. Vedovelli, and M. Voghera, *Lessico di frequenza dell’italiano parlato*, Etas Libri, Milan, 1993.

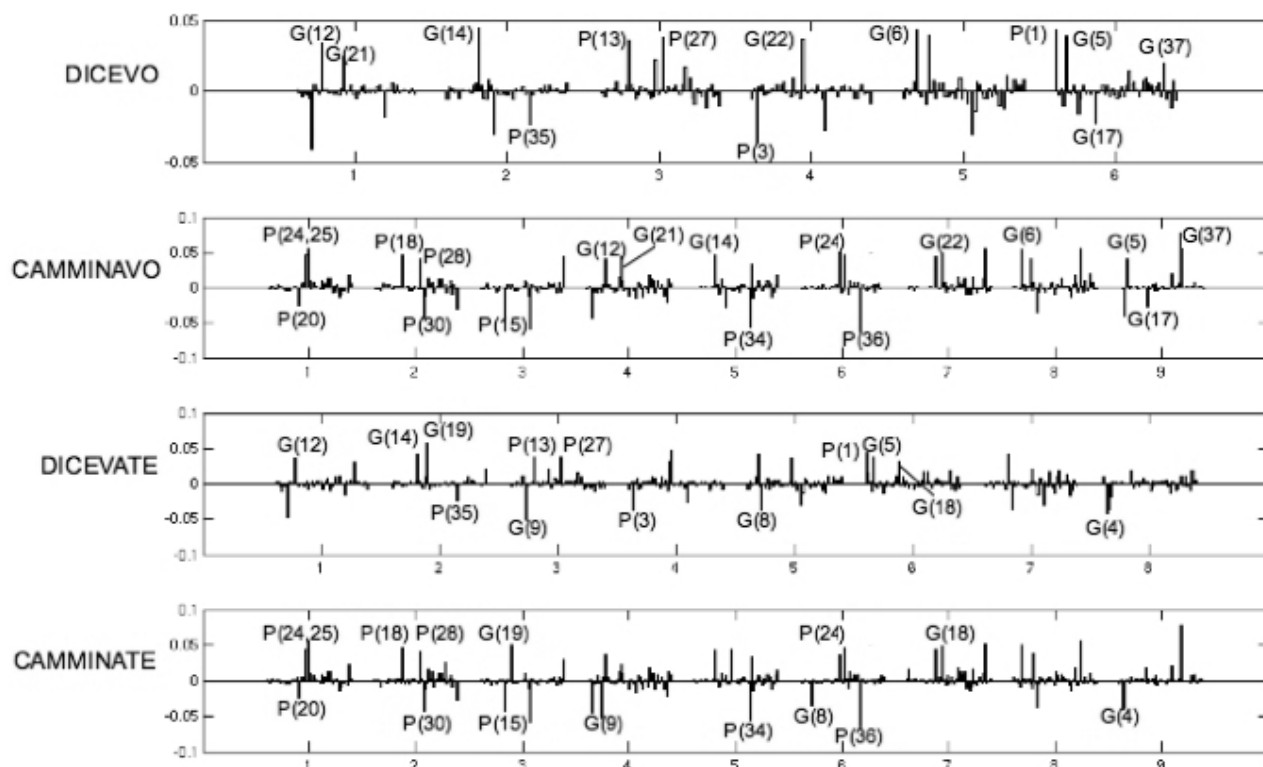


Figure 5. Emergent and distributed representations for *dicevo* ('I was saying'), *camminavo* ('I was walking'), *dicevate* ('You were saying' as 2nd person plural) and *camminate* ('You walk' as 2nd person plural). The verb forms *dicevo* and *camminavo*, both grammatically defined as IMPERFECT (tense), INDICATIVE (mood) and 1st PERSON SINGULAR (person), share the same **G**-type attributes (**G**(12), **G**(21), **G**(14), **G**(22), **G**(6), **G**(5), **G**(17) and **G**(37)). In particular the components **G**(12), **G**(14) and **G**(5) are present also in *dicevate*, thus seeming to be responsible for the grammatical classes IMPERFECT and INDICATIVE. Another grammatical feature, as for example the 2nd PERSON PLURAL is detected in the two forms *dicevate* and *camminate* (**G**(19), **G**(9), **G**(8), **G**(18) and **G**(4)). Paradigmatic relations are identified by **P**-type attributes. Forms belonging to the same paradigmatic structure as *camminavo* – *camminate* (and *dicevo* – *dicevate*) are marked by **P**-type attributes indicating relations among the verbs of the same paradigmatic family. As an example we mention the following: **P**(21), **P**(24) and **P**(25) for *camminavo* – *camminate* and **P**(13), **P**(27) and **P**(3) for *dicevo* – *dicevate*. Note that positive/negative values are irrelevant for our analysis.

[24] J. Himberg and A. Hyvärinen, "Independent component analysis for binary data: An experimental study," in *Proceedings of 3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, USA, 2001, pp. 552–556.

[25] A. Albright, "Islands of reliability for regular morphology: Evidence from Italian," *Language*, vol. 78, pp. 684–709, 2002.